Please cite this article in press as: Assié et al., SNP Arrays in Heterogeneous Tissue: Highly Accurate Collection of Both Germline and Somatic Genetic Information from Unpaired Single Tumor..., The American Journal of Human Genetics (2008), doi:10.1016/j.ajhg.2008.01.012

ARTICLE

SNP Arrays in Heterogeneous Tissue: Highly Accurate Collection of Both Germline and Somatic Genetic Information from Unpaired Single Tumor Samples

Guillaume Assié,[1] Thomas LaFramboise,[1,3] Petra Platzer,[1] Jérôme Bertherat,[5] Constantine A. Stratakis,[6] and Charis Eng[1,2,3,4,*]

SNP arrays provide reliable genotypes and can detect chromosomal aberrations at a high resolution. However, tissue heterogeneity is currently a major limitation for somatic tissue analysis. We have developed *SOMATICs*, an original program for accurate analysis of heterogeneous tissue samples. Fifty-four samples (42 tumors and 12 normal tissues) were processed through Illumina Beadarrays and then analyzed with *SOMATICs*. We demonstrate that tissue heterogeneity-related limitations not only can be overcome but can also be turned into an advantage. First, admixture of normal cells with tumor can be used as an internal reference, thereby enabling highly sensitive detection of somatic deletions without having corresponding normal tissue. Second, the presence of normal cells allows for discrimination of somatic from germline aberrations, and the proportion of cells in the tissue sample that are harboring the somatic events can be assessed. Third, relatively early versus late somatic events can also be distinguished, assuming that late events occur only in subsets of cancer cells. Finally, admixture by normal cells allows inference of germline genotypes from a cancer sample. All this information can be obtained from any cancer sample containing a proportion of 40–75% of cancer cells. *SOMATICs* is a ready-to-use open-source program that integrates all of these features into a simple format, comprehensively describing each chromosomal event.

Introduction

SNP arrays simultaneously and reliably genotype hundreds of thousands of single nucleotide polymorphisms (SNPs), and they have moved genetic studies of tumor samples into the high-throughput era (see Web Resources below). SNP arrays are powerful for the identification of chromosomal aberrations, especially deletions or amplifications.[1] The power in tumor analysis results from the high SNP density. Each SNP provides two valuable measures for that purpose: the SNP signal intensity and the allelic imbalance. The arrays produce signal intensity measurements I_A and I_B, corresponding to the two SNP alleles A and B. The SNP signal intensity is the sum of the signal of the two alleles (I_A+I_B) and reflects the number of DNA copies. The allelic-imbalance detection relies on SNPs that are heterozygous (AB) in the germline: some of them lose allele A (genotype B0) or allele B (genotype A0) in the tumor, which indicates loss of heterozygosity (LOH). LOH is qualitative, but the allelic imbalance can be made quantitative by considering a ratio of I_B/I_A. This ratio provides precision, such as in the case of duplication of one allele: a heterozygous SNP can be identified with allelic ratios of either 1:2 or 2:1, depending on which allele is duplicated.

Interrogating multiple tumors with SNP arrays results in large datasets that require computational assistance for performance of the analyses on hundreds of thousands of SNPs. Algorithms have been designed for either the Affymetrix or Illumina platforms.[2–8] These programs differ in their sensitivity for detecting alterations, requirements for corresponding normal tissue samples, types of alterations detected, interface with other programs, and accessibility.[9–28] In general, published studies show a relatively low resolution compared to the potential of the arrays.[9–28]

One of the main limiting factors of SNP-array analysis in tumors is tissue heterogeneity. Indeed, in a tumor, cancer cells are admixed with normal cells, which dilute the somatic cancer cell information.[29] Equally importantly, tumor stromal cells can contain both genomic and epigenomic alterations, often distinct from those in the epithelial neoplasia.[30–32] Microdissection can improve this issue.[16,17,21,25] However, absolute and complete separation of the cancer cells from the normal or other "contaminating" cells is quite challenging. Several investigators propose the use of immortalized cell lines, but other problems could ensue, such as the induction of new abnormalities introduced by the culture methods or process.[10,11,22,26,29,33,34] Current algorithms for tumor analysis either ignore or try to correct for the presence of normal cells. In a recent publication, Yamamoto et al. proposed a method with the signal arising from normal cells as an element of the normalization process.[8]

In this report, we describe a set of methods, applied to Illumina Beadarrays, that we developed to overcome tissue heterogeneity and even turn it into an advantage. We show how the normal cell contamination can be used as an internal reference to yield highly sensitive detection of somatic deletions. In addition, this internal reference

[1]Genomic Medicine Institute, Lerner Research Institute, [2]Taussig Cancer Institute, Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA; [3]Department of Genetics, [4]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA; [5]INSERM U567, Institut Cochin and Département d'Endocrinologie, Hôpital Cochin, 75014 Paris, France; [6]Section of Endocrinology and Genetics, Program on Developmental Endocrinology and Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: engc@ccf.org
DOI 10.1016/j.ajhg.2008.01.012.

helps in determining the proportion of cells in a sample harboring somatic deletions and amplifications and in "calling" both germline and somatic (tumor) genotypes. This method, which we called *SOMATICs,* compared favorably with existing programs in terms of its sensitivity and specificity when we applied in an original set of 54 samples.

Material and Methods

SNP Arrays and Tumor Samples

Fifty-four samples (42 tumors, 12 normal tissues) comprising gastrointestinal stromal tumors (GIST), paraganglioma, and pulmonary chondroma tumors from 26 patients were used. All tumors were examined histologically, and so percentage tumor and admixed normal cells are known by C.A.S. but kept blinded from G.A. and C.E. Tumor samples were snap-frozen in liquid nitrogen after surgery, then kept at −80°C and pulverized under liquid nitrogen. Genomic DNA was extracted with QIAamp DNA micro (QIAGEN, Valancia, CA) according to the manufacturer's recommendations. DNA was quantified with Nanodrop (Nanodrop Technologies, Wilmington, DE), and the quality was verified by agarose gel. SNP arrays were processed according to the manufacturer's recommendations with the Infinium II assay on HumanHap300 V2 arrays and run on the Illumina Beadstation (Genomics Core Facility, Cleveland Clinic Foundation). This study has been approved by the respective institutional review boards for the protection of human subjects at the participating institutions.

Detection of Hemizygous Somatic Deletions in Heterogeneous Tumor Samples

Somatic deletions are events occurring in cancer cells but not in normal cells. When processing a cancer-tissue sample, the presence of some proportion of normal cells is hard to avoid. Therefore, somatic deletions only occur in a subset of cells in a heterogeneous tumor sample, as compared to germline deletions that are present in all the cells. In the case of hemizygous deletions (loss of one allele), the presence of admixed normal cells without a deletion can impact the ability to detect somatic deletions by masking the true decrease in SNP intensity in the somatic cancer cells. Similarly, the identification of LOH in these heterogeneous tumor samples is often compromised due to the masking of the hemizygous (A0 or B0) SNPs in tumors by the heterozygous (AB) SNPs in "contaminating" normal cells.

The Illumina Beadstudio program provides a B-allele frequency (BAF) measurement, which is able to address this latter situation. BAF is a normalized metric that reflects the proportion of B-alleles in each SNP, e.g., 0 for an AA SNP, 0.5 for an AB SNP, and 1 for a BB SNP. When plotting the BAF of many consecutive SNPs in a normal tissue, three distinct bands appear, corresponding to the three genotypes: AA, AB, and BB (Figure S2, available online). BAF is a quantitative measurement, reflecting allelic imbalance instead of just LOH. In the case of a somatic deletion in a tissue sample, the BAF measured is a combination of cancer cell BAF, representing cancer-related deletion, and BAF from "contaminating" normal cells that do not harbor the deletion. In this situation, BAF of SNPs that are heterozygous in germline generate *abnormal* values between 0 and 0.5 (deletion of allele B) or 0.5 and 1 (deletion of allele A, Figure 1B). As a result, when plotting the BAF of many consecutive SNPs belonging to a somatic deletion in a heterogeneous sample, the unique band of heterozygous SNPs is replaced by a two-band pattern of SNPs that are heterozygous in germline and hemizygous in tumor (Figures 1B–1E). These two-band patterns actually reflect the allelic imbalance of the heterozygous SNPs associated with the deletion of one allele.

SOMATICs was developed to automatically and accurately detect somatic deletions on the basis of these specific two-band patterns. The general flowchart of *SOMATICs* is presented in Figure S1. The following four steps are applied:

(1) Identification of Abnormal SNPs

Abnormal SNPs are defined as those that are not normal AA, AB, or BB. In terms of BAF, these SNPs have BAF values significantly different from the BAF of SNPs AA, AB, and BB. For each SNP chip, *SOMATICs* determines what these normal values are by determining *reference* distributions for each of the three genotypes (see Appendix). With these distributions, a probability of being AA, AB, or BB can be determined for each SNP, and *abnormal* SNPs are identified with different BAF threshold values, which are detailed in the Appendix.

(2) Identification of the Two-Band Patterns among the Abnormal SNPs

Two-band patterns are deviations from the unique central band of normal heterozygous SNPs (BAF = 0.5). The distance of deviation can vary: the two bands can be obvious (Figure 1B), very close to the bands of homozygous SNPs (Figure 1C), or very close to each other (Figure 1E). *SOMATICs* uses three different methods to identify these SNP patterns in each of these situations (for details, see Appendices).

(3) Determination of Boundaries for each Two-Band Pattern

This step converts individual SNPs into chromosomal regions. The smoothing process (Adaptative Weights Smoothing[35]) identifies all the consecutive SNPs that should be considered as having a constant BAF (for details, see Appendices).

(4) Calling Deletions in the Smoothed Fragments

For each SNP, Illumina Beadstudio generates the logR ratio metric, which reflects the number of DNA copies. The logR ratio is a log-transformed ratio of the measured SNP signal intensity by the expected intensity if two copies of DNA are present. This logR ratio is normalized so that two copies of DNA generate a logR ratio ≈0, whereas one copy of DNA generates a logR ratio ≈−0.5. Use of the logR ratio to call a deletion might seem the most straightforward. However, the logR ratio is less responsive than the BAF,[29] and in the case of somatic hemizygous deletions, this difference is even more important (see Results and Discussion section). For this reason, instead of just considering the logR ratio, *SOMATICs* uses the BAF to detect all potential deletions and subsequently uses the logR ratio to confirm the deletion.

To call a region "deleted," the following three criteria must be met:

i. There must be a significant decrease of the logR ratio.
ii. There must be a shift of the logR ratio that is concordant with the proportion of cells harboring the deletion. For instance, for a deletion occurring in almost all the cells in a sample, the logR ratio is expected to be close to −0.5, whereas for a deletion occurring in very few cells, the logR ratio is expected to be close to 0.
iii. The boundaries of the two-band pattern and the region with decreased logR should be concordant (for details, see Appendices).

Detection of Other Chromosomal Aberrations in Heterogeneous Tumor Tissue Samples

Germline Deletions

Unlike somatic deletions, germline deletions of one allele are chromosomal aberrations occurring in all the cells of an individual.

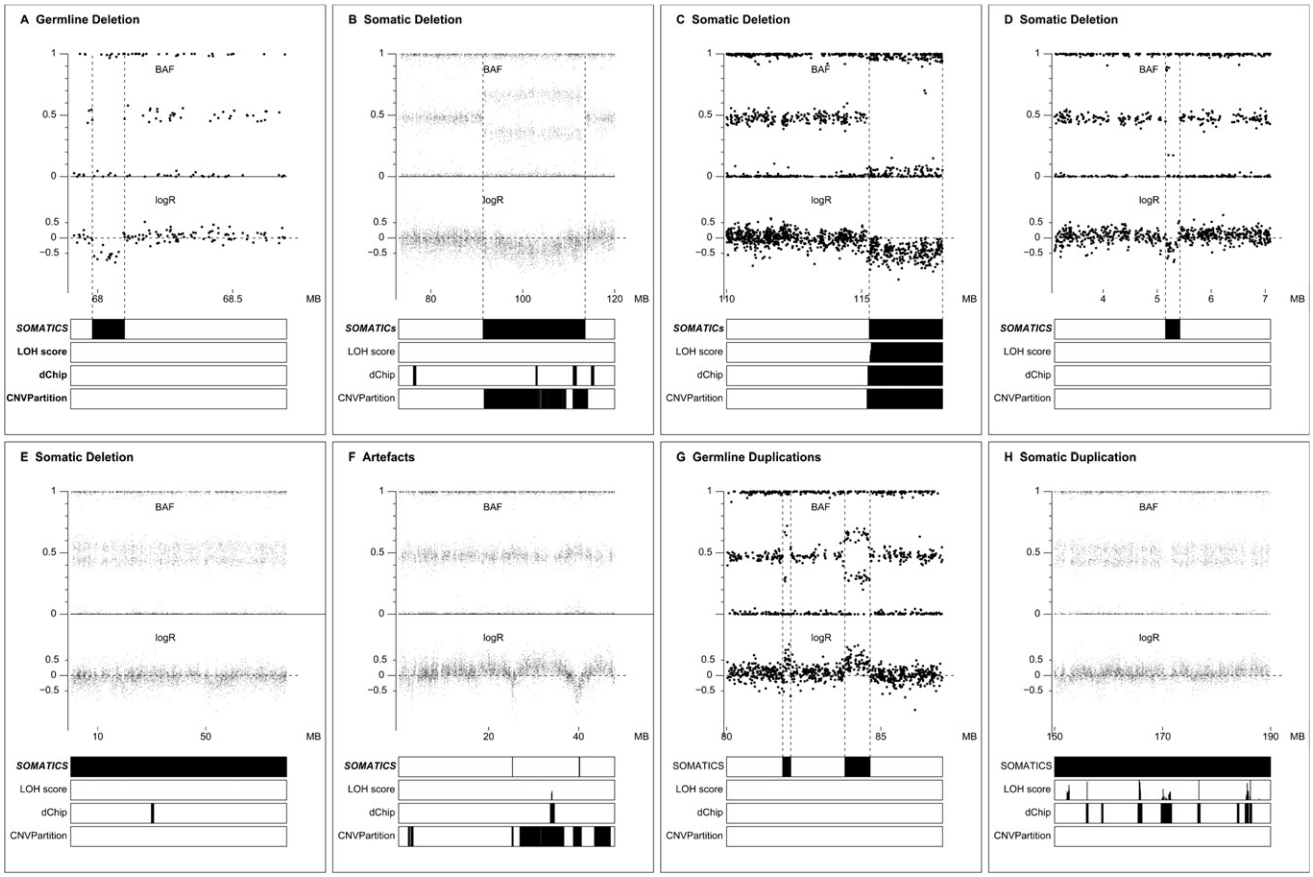

Figure 1. Automatic Detection of Chromosomal Aberrations in Heterogeneous Tissue Samples

In each panel, the BAF and the logR ratio are plotted along chromosomal regions. The black boxes at the bottom represent detection with *SOMATICs* and other currently available methods, namely, Beadstudio LOH score (y axis range from 0 to 5), dChip (default detection threshold), and CNVPartition.

(A) Small germline deletion is revealed as a decreased logR ratio (−0.5), and the band of heterozygous SNPs centered on 0.5 is absent on the BAF plot.

(B–E) Various types of somatic deletions revealed on the BAF plot in which the single band of heterozygous SNPs is replaced by two bands and the logR ratio is decreased, but to a lesser extent as compared to germline deletions. The various types of somatic deletions are manifested by differences in the position and the size of the two-band patterns. Note that (E) shows a somatic deletion occurring in very few cells. In this situation, detection is easier with use of the BAF (two-band pattern) than with the logR (reduced shift downward). However, the copy number call as a "deletion" relies on a significant decrease of the logR (in this situation, $p < 2.2e^{-16}$).

(F) Wavy fluctuations of logR ratio, which is not reflected in the BAF, are artefacts. This artifact is responsible for false-positive detection by programs focusing only on logR ratio.

(G) Two small germline duplications are revealed on the BAF plot as heterozygous SNPs showing a two-band pattern with a logR ratio that is increased.

(H) In somatic duplication, the two bands are closer to one another and the increase in logR is less than that of germline amplification (G). As with (E), for (H), the BAF two-band pattern is easier to detect than is the logR shift upward. However, the copy number call as an "amplification" relies on a significant increase of the logR (in this case, $p < 2.2e^{-16}$). *SOMATICs* can detect and differentiate the various types of alterations when other programs cannot.

Therefore, hemizygous germline deletions can be detected in heterogeneous tumor samples via hemizygosity, whereby hemizygous regions generate homozygous SNP genotype calls and decreased signal intensity. *SOMATICs* identifies a germline deletion as any consecutive stretch of ≥ 5 SNPs, identified with BAF values of 0 (A0) or 1 (B0), which also have a decreased logR intensity. The logR decrease is ascertained as described previously with the three criteria detailed in (4) ("Calling Deletions in the Smoothed Fragments"):When these criteria are not met, long stretches of homozygous SNPs can reveal uniparental disomy. The discrimination between uniparental-disomy-related stretches and randomly occurring stretches of homozygous SNPs can be performed with existing programs such as dChip.[5] When two alleles are deleted (homozygous deletions), the logR of SNPs dramatically drops to very low values (< -1); *SOMATICs* calls two or more consecutive SNPs with such logR values as homozygous deletions. See Figure 1A.

Somatic and Germline Amplifications

Similar to somatic deletions, amplifications of one allele result in allelic imbalance of heterozygous SNPs, which can be detected by BAF two-band patterns, as described previously. This occurs in both germline (Figure 1G) and somatic amplifications

(Figure 1H), and thus, *SOMATICs* can detect both with the same strategy as that for somatic deletions (except that the logR ratio is now increased): when two bands are close to each other in the BAF two-band pattern and the logR ratio shift is decreased (Figure 1H), a somatic amplification is detected.

Proportion of Cells Harboring Somatic Deletions and Amplifications in Heterogeneous Tissue Samples

In heterogeneous tissue samples, the allelic imbalance related to somatic deletions and amplifications of one allele can be identified by the BAF two-band patterns described previously. The position of the two bands is determined by the relative proportions of cells harboring a given somatic event versus cells without this event.[10] For example, in the case of a somatic deletion, the two bands are close to the bands of homozygous SNPs when the tissue contains a majority of cells harboring the deletion (Figure 1C). In contrast, when the tissue contains very few cells harboring the deletion, the two bands are close to each other in the center of the BAF plot (Figure 1E). On the basis of this feature, *SOMATICs* provides an assessment of the proportion of cells, "*c*," harboring a somatic event. We introduce *c* in the three most common somatic events: hemizygous deletions, duplication of one allele, and deletion of one allele with duplication of the other. For each situation, the exact relationship between *c* and the allelic ratio I_B/I_A can be specified as follows:

Somatic deletion of allele B:

$$I_B/I_A = 1 - c \qquad \text{[Equation 1]}$$

Somatic duplication of allele A:

$$I_B/I_A = 1/(1 + c) \qquad \text{[Equation 2]}$$

Somatic deletion of allele B and duplication of allele A:

$$I_B/I_A = (1 - c)/(1 + c) \qquad \text{[Equation 3]}$$

In addition, Illumina defines BAF as:[10]

$$\tan(\mathrm{BAF}) = I_B/I_A \qquad \text{[Equation 4]}$$

Combining equations 1, 2, and 3 with equation 4, we can explicitly solve for *c* as a function of BAF:

Somatic deletion of allele B:

$$c = 1 - \tan(\mathrm{BAF}_{\mathrm{del}} \times \pi/2) \qquad \text{[Equation 5]}$$

Somatic duplication of allele A:

$$c = 1/\tan(\mathrm{BAF}_{\mathrm{ampl}} \times \pi/2) - 1 \qquad \text{[Equation 6]}$$

Somatic deletion of allele B and duplication of allele A:

$$c = \left[1 - \tan(\mathrm{BAF}_{\mathrm{deldupl}} \times \pi/2)\right]/\left[1 + \tan(\mathrm{BAF}_{\mathrm{deldupl}} \times \pi/2)\right] \qquad \text{[Equation 7]}$$

Among these three formulae, the appropriate one is chosen according to the logR ratio. In addition, these formulae require only one BAF value, yet there are two BAF values for each two-band pattern. Therefore, the two bands are transformed into a single band by "folding" the BAF plot so that the upper band is superimposed onto the lower band (Figure 2A). The median value of the SNPs in the unique band is used for computing *c*(for details, see Appendices).

Germline Genotype Inference from Cancer Samples

In normal tissues, BAF plots show three clearly distinct bands of SNPs, each corresponding to one genotype (AA, AB, or BB). In cancer samples, the majority of chromosomal regions do not harbor any aberrations. In this case, genotype inference is clearly discernable with the use of fixed BAF thresholds (Figure 3A). In chromosomal regions with somatic deletions or amplifications, the unique band of heterozygous SNPs is replaced by two bands: they correspond to SNPs that are heterozygous in the germline. *SOMATICs* systematically looks for these two-band patterns: the SNPs that belong to the two bands are called heterozygous in germline. This genotyping is highly reliable as far as these two bands can be discriminated from the homozygous SNPs. This is true for the vast majority of chromosomal aberrations in a majority of cancer samples. However, in the specific case of a hemizygous deletion occurring in almost all the cells of a sample, the two bands are difficult to discriminate from the homozygous SNP bands (Figure 1C). In this case, *SOMATICs* uses a strategy of best discrimination to warrant the accuracy of the genotype (for details, see Appendices).

Third Party Softwares

All original scripts were written in R, version 2.4.[36] All the scripts are available as supplemental data. Specific R packages dip test and Adaptive Weights Smoothing (aws) were included in *SOMATICs*.[37,38] BeadStudio 3.1 was used to generate the BAF and logR ratios, which were exported as text files. The LOH score, Chromozone, and CNVPartition plug-ins provided with this version of BeadStudio were used to identify LOH, allelic imbalance, and CNV. dChip[7] (release Dec 7, 2006) was applied to the normalized allele intensities X and Y obtained from Illumina BeadStudio 3.1, along with the genotype calls, as recommended. The SNP annotations were also included (physical positions provided by Illumina). The LOH analysis was performed with Hidden Markov Models for unpaired data, assuming a proportion of heterozygous SNPs of 35% for the Illumina HumanHap300 (determined from the normal samples). All other parameters were set to default values.

Results

Accurate Detection of Deletions and Amplifications in Heterogeneous Tissue Samples with SOMATICs

SOMATICs automatically detects the deletions and amplifications in a tissue sample processed on Illumina Beadarrays. The originality of the approach is that it can specifically interpret the somatic events occurring in a subset of cells within a sample, typically in the cancer cells of a heterogeneous tissue sample "contaminated" by normal cells. In the case of one allele's somatic deletions in particular, the presence of normal cells is responsible for specific two-band patterns of allelic imbalance with the Illumina BAF measurement. Similar two-band patterns are associated with the amplification of one allele, another case of allelic imbalance. *SOMATICs* automatically detects these two-band patterns (Figure 1), even when somatic deletions or amplifications occur in a small proportion of the cells (Figures 1E and 1H) or when the somatic events are physically small (Figures 1D and 1G).

Compared to other algorithms, *SOMATICs* deals better with tissue heterogeneity. For example, methods of looking for deletions based on LOH detection (dCHIP[5] and

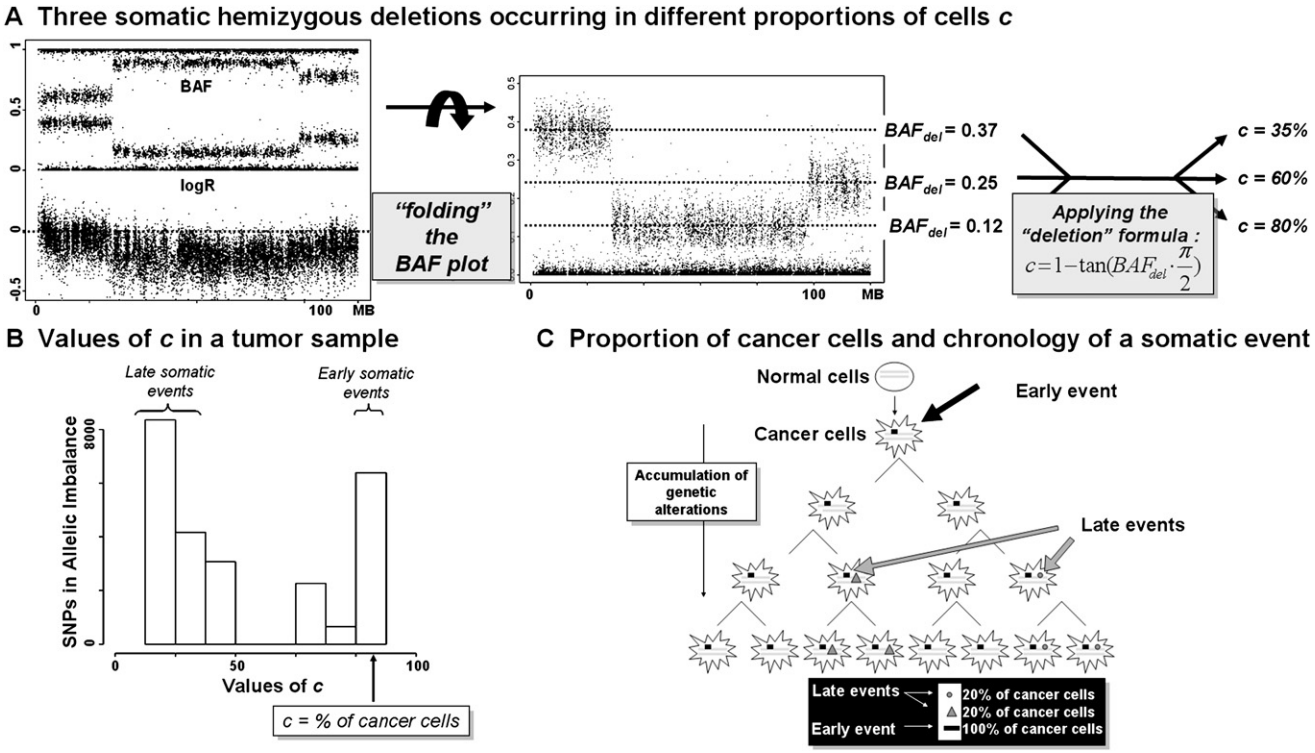

Figure 2. Proportion of Cells, "*c*," in a Heterogeneous Tumor Sample Harboring a Somatic Genetic Event

(A) BAF and the logR ratio plots from one chromosome reveal three somatic hemizygous deletions occurring in three different proportions of cells. The BAF_{del} of heterozygote SNPs is measured after "folding" the BAF plot along an axis centered around 0.5. Then BAF_{del} is used to determine *c* with the formula designed for somatic deletions (see Appendices for details).

(B) Frequency distribution showing the number of SNPs included in the somatic deletions by the proportion of cells, "*c*," in which these events occur. Some somatic deletions occur in over 80% of cells (rightmost bar). Assuming that only cancer cells harbor somatic deletions, the proportion of cancer cells is then estimated as 80% in this sample.

(C) Schematic illustrating the relationship between the chronology of somatic events during tumorigenesis and the proportion of cancer cells with these events. Early somatic events are present in all (or a great majority of) cancer cells, whereas late somatic events are only present in subsets of cells.

LOH score, Illumina Beadstudio) detect only somatic deletions occurring in almost all the cells of the sample (Figure 1C). These methods cannot detect deletions occurring in lower proportions of cells (Figures 1B and 1E), germline amplifications (Figure 1G), or somatic amplifications (Figure 1H). In addition, LOH-detection approaches cannot reveal small chromosomal events (Figure 1D). Like *SOMATICs*, Chromozone (Illumina Beadstudio) detects the BAF two-band patterns indicating allelic imbalance. However, the sensitivity of Chromozone is limited: small deletions (Figure 1D) or deletions occurring in very few cells (Figure 1E) are not detected. In addition, Chromozone provides only graphic bookmarks, without tabular output or boundary definitions. CNVPartition (Illumina) is solely based on the logR ratio. However, in the case of somatic events, normal cell contamination dilutes the logR ratio, decreasing its sensitivity (Figures 1B–1H). Note that the human eye can barely perceive the difference between plots in Figure 1E (downward for deletion) and Figure 1H (upward for amplification). However, *SOMATICs* detects this shift of logR upwards (amplification) or downwards (deletion) with a $p < 2.2e^{-16}$. In addition, CNVPartition generates false positives in noisy samples, unlike *SOMATICs* (Figure 1F).

With *SOMATICs*, germline deletions can be distinguished from somatic deletions, even when the latter occur in almost all the cells of a tissue sample. This discrimination is based on the identification of the specific BAF two-band pattern that is associated with somatic deletions but not with germline deletions. Figure S3 shows two regions of the same size and SNP density, one with a somatic deletion occurring in almost all of the cells (two-band pattern very close to the homozygous SNPs), the other with a germline deletion (without the two-band pattern). *SOMATICs* discriminates between these two situations.

Unique integrated outputs are also generated by *SOMATICs*. All relevant information associated with each chromosomal aberration is gathered, namely the boundaries, the aberration type (deletion, amplification, deletion and/or duplication), the affected compartment (germline or somatic), the proportion of cells in the sample that harbor the aberration (see next paragraph), and the criteria reflecting the reliability of the finding (significance of statistical tests and size of the fragment).

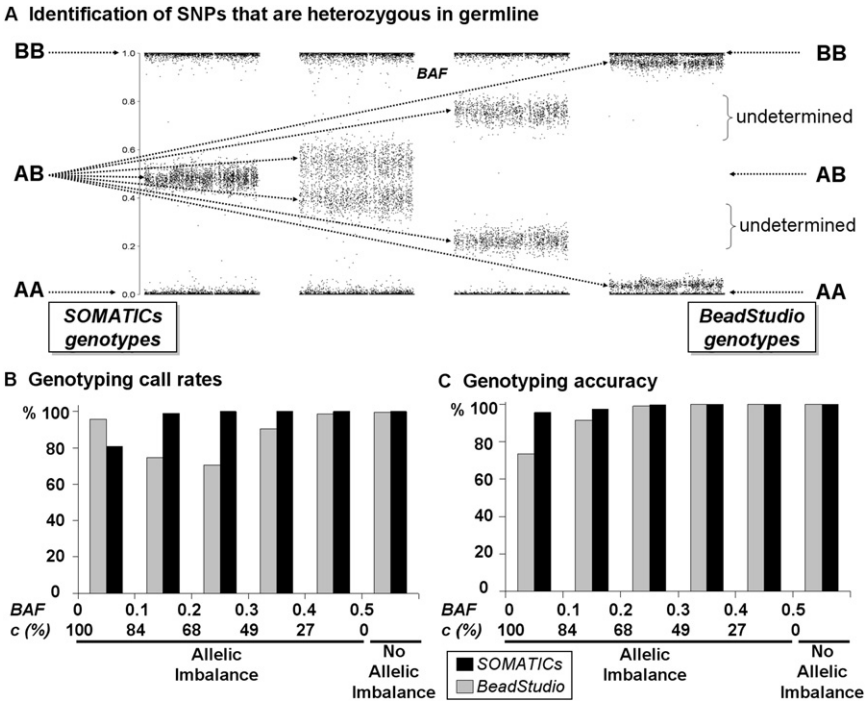

Figure 3. Germline Genotype Inference from a Heterogeneous Cancer Sample
(A) Schematic representation of the calling methods used by *SOMATICs* and by BeadStudio. *SOMATICs* is specifically designed to generate appropriate calls in genomic regions with allelic imbalance. When bands of heterozygous AB SNPs can be discriminated from the bands of homozygous AA and BB SNPs, a reliable germline genotype call can be provided.
(B) illustrates enotyping call rates in regions of allelic imbalance with *SOMATICs* and BeadStudio, and (C) illlustrates genotyping accuracy in regions of allelic imbalance with *SOMATICs* and BeadStudio. These results were obtained by comparison of genotypes inferred from tumors with genotypes read in normal corresponding tissue obtained with seven pairs of matched samples. The results in (B) and (C) are expressed as a function of the distance from the heterozygote bands to the homozygote bands, on a scale of 0 to 0.5, representing the mean BAF of the lower band of heterozygote SNPs. This mean BAF is converted into a proportion of cells, "*c*," with allelic imbalance by use of the formula for somatic deletions (see Appendices for details).

Proportion of Cells with a Somatic Event

In heterogeneous tissue samples, somatic deletions and amplifications occur only in a subset of cells. *SOMATICs* is able to determine the proportion of cells, "*c*," in the sample that harbor the somatic event (Figure 2B). For validating the whole procedure for estimating *c*, data from a serial dilution of cancer cells with normal matched cells generated by Pfeiffer et al.[10] were used. These experiments included SNP-array runs for each dilution of cancer cell proportions of 0, 25, 50, 75, and 100%. A complex set of consecutive chromosome aberrations that included a large deletion of one allele harbored by all of the cancer cells (13q distal) was shown by the investigators. In the chromosomal region with the deletion, the median BAFs of the lower band on the two-band patterns were at approximately 0.5, 0.45, 0.35, 0.15, and 0 for the different dilutions. Using these rough values with our "deletion" formula within *SOMATICs*, we find that the estimated proportion of cancer cells, "*c*," are 0, 15, 39, 76, and 100%, which are close to the actual dilutions at 0, 25, 50, 75, and 100%.

Determination of the Tumor Content

When the proportions of cells determined for all the somatic deletions in a sample are gathered, a distribution of these values can be generated. Figure 2B shows an example with two peaks in the distribution. The highest peak is centered at 80%. That means that some somatic deletions occurred in up to 80% of the cells of the sample. If we assume that only cancer cells, not normal cells, harbor deletions, then 80% of the cells in the sample are cancer cells. Tumor content is difficult to measure, even with anatomopathological approaches. To our knowledge, this is the first computational method that determines the tumor content.

A first application of the proportion of cancer cells is obvious: when the tumor content is very low, one would expect a high false negative rate for identifying somatic events. Therefore, the sensitivity of an SNP-array experiment to detect somatic events can now be assessed.

Germline Genotype Inference

SOMATICs is uniquely able to infer the germline genotype from a cancer sample, even in regions with allelic imbalance (Figure 3). To assess the reliability of the germline calls made by *SOMATICs*, we compared the germline calls inferred from 17 tumors with the genotype calls generated from their normal corresponding tissue. With the use of *SOMATICs* in the cancer samples, 99.6% of the genotypes were called, and among them, the genotyping accuracy was 99.75%.

To assess the impact of allelic imbalance on the germline genotyping, we measured the genotyping call rates and accuracy specifically in chromosomal regions with somatic

deletions or amplifications. When the bands of the two-band patterns are distinct from the homozygous SNPs on the BAF plot, call rates remain high and genotyping errors are low (Figure 3A). When the BAF of the lower band in a two-band pattern is ≥ 0.2, the call rate is 99.9% and the genotyping accuracy is higher than 99.6%. This limit corresponds to a somatic deletion occurring in $\approx 70\%$ of the cells in a sample (Figure 3B and 3C). When the bands come closer to the bands of homozygous SNPs (e.g., BAF of the lower band between 0.1 and 0.2 [representing a somatic deletion in 84% of the cells in a sample]), the call rate remains at 98.8%, with a genotyping accuracy decreasing to 97.6%. For comparison, in this situation, BeadStudio germline genotyping is associated with a low call rate (74.5%) and low genotyping accuracy (91.5%). When the bands come even closer (BAF of the lower band < 0.1), discrimination between germline heterozygous and homozygous SNPs become challenging. In this situation, *SOMATICs* achieves a call rate of 80.5%, at an accuracy of 95.5%. For comparison, the call rates from BeadStudio are high (95.6%), but, importantly, this is due to the incorrect calling of germline heterozygous SNPs as homozygous, therefore providing an accuracy of 73.6%.

Discussion

Up until now, SNP-based analysis for somatic deletion or amplification required paired germline and tumor samples. Furthermore, cellular heterogeneity in tumors, usually normal and tumor cellular admixture, has proven problematic for such SNP-based studies. *SOMATICs* was designed for use in heterogeneous tissue studies, such as those utilizing tumor tissue samples. We can also imagine that such a technique might be amenable to examination for genetic alterations that occur in a germline mosaic manner. The main innovation of *SOMATICs* is the utilization of normal cell admixture as an advantage rather than a liability. *SOMATICs* enables the automatic detection of deletions and amplifications from Illumina Beadarrays, a feature already proposed by other methods.[5,7,29] However, compared to these other available methods, *SOMATICs* appears to be quite sensitive, even when other methods fail to detect very small deletions or amplification. Both allelic imbalance and DNA copy-number alterations related to deletions and amplifications are integrated into a single analysis in *SOMATICs*, and the background "noise" from each SNP array is taken into account. An added advantage of the program's design is that the program code is open and modular, permitting one to perform custom modifications; e.g., to run a specific analysis or to generate specific output formats.

Two original features are implemented in *SOMATICs*: the proportion of cells harboring a somatic event and the germline genotype inference from a cancer sample. One extrapolated utility for the ability to determine the proportion of a somatic genetic event is the ability to quickly predict relatively early and late somatic events. Because of accumulated work over the last two decades, it is generally accepted that relatively early somatic events have a higher frequency of cells harboring those events, as compared to later somatic events in which there would be a lower frequency. For example, Figure 2C is a schematic representation of the genetic accumulation of the somatic chromosomal aberrations throughout the clonal growth of the tumor. At the end, the proportion of cancer cells with a somatic event reflects the chronology of the event: early somatic events are present in a majority of cancer cells, whereas late somatic events are present only in a small subset of cancer cells. Going back to the distribution of proportions of cells harboring somatic events in Figure 2B, if we assume that 80% of the cells are cancer cells, then the events occurring in 80% of the sample cells (i.e., in the majority of the cancer cells) can be qualified as relatively earlier somatic events, whereas those occurring in 35% of the sample cells (i.e., in less than half of the cancer cells) can be qualified as relatively later somatic events. Finally, the ability to infer germline genotype from only tumor samples without the physical existence of corresponding germline samples makes this a useful method, because corresponding germline tissue might not always be attainable. *SOMATICs* is able to achieve an accuracy of $> 95\%$ in this regard, compared to $< 75\%$ for Beadstudio. This is corroborated by a completely independently performed study[39] to look for germline deletions in four selected candidate genes—SD*HB, SDHC, SDHD* and *PDGRFA*—in a proportion of samples that were common to this study. When the results were compared, the germline deletions identified by the experimental candidate-gene germline-deletion-analysis study (e.g., see Figure 7 of Pasini et al.[39]) were noted as germline-deletion calls by *SOMATICs*.

Appendix A. Detection of Somatic Deletions in Heterogeneous Tumor Samples

The detection of somatic deletions is performed by these four steps (Figure S1):

I. Identification of Abnormal SNPs that are neither Entirely Homozygous nor Entirely Heterozygous on the basis of BAF

To ascertain that a SNP is neither a "normal" homozygous SNP nor a "normal" heterozygous SNP on the basis of its BAF, one needs to first determine the BAF values of homozygous and heterozygous SNPs that are "normal," e.g., not belonging to any amplification or deletion event. Distributions of normal values are determined separately for each array, because these distributions will reflect the experimental variability from array to array. Toward this end, *SOMATICs* looks for "normal" chromosomes for each sample. Such chromosomes can be identified as having three thin bands on the BAF plot (Figure S2), each corresponding to one of the three genotypes (AA, AB, and BB). A

temporary assignment of SNPs to one of the three genotypes is performed with BAF thresholds of 0.25 and 0.75 (Figure S2). Assuming that there is always at least one normal chromosome in any sample, this chromosome is selected as the one with the lowest BAF variance for each of the three genotypes. The BAF values for the SNPs from this chromosome are used as *reference* distributions for each of the three genotypes.

These reference distributions are used to define three BAF regions, termed "blue," "green," and "red," displayed in Figure S2. A separate algorithm is used for each of these regions to identify *abnormal* SNPs and their organization into two-band patterns:

1. In the blue regions, *abnormal* SNPs are close but distinct from homozygous SNPs. The BAF values of these *abnormal* SNPs are defined to be between 0.25 and the 95th percentile of the AA *reference* distribution, or between 0.75 and the 95th percentile of the BB *reference* distribution.
2. In the green regions, *abnormal* SNPs are obviously abnormal, given that they are distinct from the *reference* AA, AB, and BB regions. The BAF values of these *abnormal* SNPs are defined to be between the 99.9th percentile of the AA reference distribution and the 99.9th percentile of the BB reference distribution.
3. In the red region, *abnormal* SNPs have BAF values close to 0.5. This region is defined to be between 0.25 and 0.75.

II. Identification of BAF Two-Band Signature of Allelic Imbalance among the Abnormal SNPs

Three specific algorithms are applied to detect the BAF two-band patterns of allelic imbalance among the *abnormal* SNPs:

1. In the green regions, the two-band patterns are obvious and distant from the homozygous SNP bands (Figure 1B and Figure S2B). In this case, the two-band patterns are defined as any succession of $\geq$ three SNPs in the green region. These SNPs can be consecutive, or they can be separated by homozygous SNPs, given that hemizygous SNPs are called "homozygous" with the BAF.
2. In the blue regions, the two-band patterns display bands of SNPs close to the homozygous SNPs (Figure 1C and Figure S2). In this case, the two-band patterns are identified with a likelihood ratio, which divides the likelihood that the *abnormal* SNPs are heterozygous in the germline by the likelihood that the *abnormal* SNPs are homozygous. The likelihood that the *abnormal* SNPs are heterozygous in the germline is estimated by the concordance of the relative proportion of heterozygous and homozygous SNPs with the expected 1/3 proportion of heterozygous SNPs observed for the entire SNP array, by use of a Chi-square test (or a Fisher's exact test when there are less than five heterozygous or homozygous SNPs or less than 20 SNPs). The likelihood that the *abnormal* SNPs are homozygous is computed as the product of the probabilities that each *abnormal* SNP is actually a homozygote. These probabilities are determined with the probability-distribution function of the *reference* AA and BB distributions [see section i above]. Any succession of $\geq$ five *abnormal* SNPs with BAFs close to 0 or 1, e.g., belonging to the blue regions, [see section i above] is tested. With this approach, subchromosomal regions with long stretches of homozygous SNPs generate likelihood ratios lower than 20, and subchromosomal regions with bands of heterozygous SNPs overlapping with the homozygous SNPs generate likelihood ratios over 100 (Figure S3). A decision threshold of 40 was chosen.
3. In the red regions, the two-band pattern displays two bands close to each other (Figures 1E and 1H, Figure S2). In this case, the two-band patterns are defined as regions with a bimodal BAF distribution of heterozygous SNPs, in contrast to normal heterozygous SNPs that show a unimodal distribution. The mode of the distribution is identified by application of the Hartigan's dip test for unimodality[40] to all of the heterozygous SNPs with BAFs between 0.25 and 0.75. Moving windows of three widths are used (1000, 200, and 50 SNPs), with steps of 1/25th the width. In each window, a dip score is generated. The dip score is a nonparametric statistic that measures deviation from unimodality. A null distribution for the dip score is generated via random sampling for each window size and for each sample. For instance, for the 200 SNP window, 1000 dip scores are calculated from 200 BAFs randomly sampled from the *reference* heterozygous SNPs, generating the distribution of dip scores for heterozygous SNPs *without* allelic imbalance. The 99^{th} percentile of this distribution is used as a threshold for calling allelic imbalance. All SNPs located inside a positive window $\pm$ 1/2 step are preliminarily considered to be in allelic imbalance. This strategy is very sensitive, but it might generate false-positive calls, especially for normal SNPs that are near a cluster of SNPs with an obvious bimodal distribution. In this situation, however, because the bimodal portion of the window is obvious, the segmentation procedure (see section III below) will identify the different segments. Finally, the BAFs of heterozygous SNPs in each segment are compared to the *reference* BAF distribution for heterozygote SNPs by use of either a Student t test (if $N > 20$) or a Wilcoxon test ($N \leq 20$). Only segments that reach significance (two-sided p values < 0.05) in this test are deemed to be in allelic imbalance.

III. Determination of Boundaries for Each Two-Band Allelic Imbalance Pattern

In order to convert successions of SNPs into chromosomal regions and determine the boundaries of each region in allelic imbalance, the Adaptative Weights Smoothing (AWS) procedure[36] (as implemented in the AWS R package) is applied as follows:

SNPs belonging to the two-band patterns are selected as the SNPs having BAF values between the 99th percentile of the *reference* distribution of AA SNPs and the 99th percentile of the *reference* distribution of homozygous BB SNPs. The two bands are converted into a single band by "folding" the BAF plot so that the two bands are superimposed (Figure 2A). The "folding" is centered on the mean BAF of *reference* heterozygous SNPs (instead of 0.5). Indeed, careful analyis with the *reference* heterozygous SNPs reveals that the mean BAF (denoted $BAF_{\mu het}$) of normal heterozygous SNPs is actually lower than 0.5, with a specific value for each experiment. For the "folding," any SNP with a BAF higher than $BAF_{\mu het}$ was converted to fit into the [0, $BAF_{\mu het}$] interval, by use of the following transformation:

$$BAF_{converted} = BAF_{\mu het} - (BAF - BAF_{\mu het}) \times BAF_{\mu het}/(1 - BAF_{\mu het}) \quad \text{[Equation 1]}$$

These SNPs are submitted into the ASWH function. The ASWH function is used with the following parameters: hinit = 3, hmax = 500, p = 0, sigma2 = 0.01. This procedure generates a "smoothed" value for each SNP. Consecutive SNPs with smoothed values that differ by less than 0.02 units are deemed to be part of the same segment.

IV. Testing for DNA Copy-Number Variation within the Two-Band Pattern of Allelic Imbalance

For each chromosomal segment in allelic imbalance identified as described above, the logR ratio is explored. The following three criteria have to be met to call a deletion:

1. There must be a significant decrease of the logR ratio: the mean logR ratio of the SNPs within that region is compared to a *reference* logR ratio distribution. This *reference* distribution is obtained from distribution of logR arising from a normal chromosome in the sample, defined as the one with the smallest logR variance and no aberrations as ascertained by a unique band of heterozygous SNPs on the BAF plot. The distribution is centered at 0 by subtracting its mean. The comparison between the SNPs from the region in allelic imbalance and the *reference* logR distribution is performed by use of either the Student t test (more than 20 SNPs) or the nonparametric Wilcoxon rank sum test (less than 20 SNPs). Significance is called for two-sided p values < 0.05.
2. The decrease of the logR ratio must be concordant with the proportion of cells, "*c*" harboring the somatic event. The median logR ratio of SNPs within the region in allelic imbalance must be less than $[-0.3 \times c]$.
3. The boundaries of the region in allelic imbalance and of the region with decreased logR ratio should be concordant. For this criterion, *SOMATIC*s screens the boundaries of the two-band patterns for a shift in the logR ratio. The criterion is met when the median logR ratio of the five SNPs outside each boundary is at least $[0.3 \times c]$ units higher than the median logR ratio of the SNPs within the boundaries.

Appendix B. Proportion of Cells Harboring Somatic Deletions and Amplifications in Heterogeneous Tissue Samples

BAF is a normalized measurement of the B/A allelic ratio. The non-normalized measurement is denoted "θ" in the original Illumina Beadstudio paper, with:

$$\text{Theta} = (2/\pi) \times \arctan(I_B/I_A) \quad \text{[Equation 2]}$$

where I_A and I_B are the normalized intensities of A and B. Theta is normalized into BAF by linear interpolation with "canonical" theta values obtained from a panel of normal individuals so that BAF values are approximately 0, 0.5, and 1 for SNPs with genotypes AA, AB, and BB, respectively. As a result, we have:

$$I_B/I_A = \tan(BAF \times \pi/2) \quad \text{[Equation 3]}$$

with $I_B/I_A = 0$ for BAF = 0 (genotype AA, B allele signal = 0), $I_B/I_A = +\infty$ for BAF = 1 (genotype BB, A allele signal = 0), and $I_B/I_A = 1$ for BAF = 0.5 (genotype AB, A allele signal = B allele signal).

We let "*c*" denote the proportion of cells in the sample harboring a somatic event. In the case of a somatic hemizygous deletion of allele B, the signal from allele B arises only from the normal cells present in the sample, which are in proportion $(1 - c)$. Therefore the signal from allele A is decreased by a factor of $(1 - c)$, and the I_B/I_A ratio is increased by a factor of $(1 - c)$. For heterozygous SNPs, I_B/I_A is equal to 1. Using Equation 3, we then obtain:

$$1 - c = \tan(BAF_{del} \times \pi/2) \quad \text{[Equation 4]}$$

which can be converted into

$$c = 1 - \tan(BAF_{del} \times \pi/2). \quad \text{[Equation 5]}$$

In the case of a somatic amplification via duplication of allele A, the signal from allele A is doubled in the cancer cells, which are in proportion *c*, and unchanged in the normal cells, which are in proportion $(1 - c)$. Therefore, the signal from allele A is increased by a factor of $[2 \times c + (1 - c) = (1 + c)]$. Therefore, the B/A ratio is decreased by a factor of $(1 + c)$. For heterozygous SNPs, I_B/I_A is equal to 1. Using Equation 3, we obtain:

$$1/(1 + c) = \tan(BAF_{ampl} \times \pi/2) \quad \text{[Equation 6]}$$

so that

$$c = 1/\tan\left(\mathrm{BAF}_{\mathrm{ampl}} \times \pi/2\right) - 1 \qquad \text{[Equation 7]}$$

In the case of a somatic deletion of allele B with duplication of allele A, the signal from allele A is doubled in the cancer cells, which are in proportion c and unchanged in the normal cells, which are in proportion $(1 - c)$. Therefore, the signal from allele A is increased by a factor of $[2 \times c + (1 - c) = (1 + c)]$. The signal from allele B arises only from the normal cells present in the sample, which are in proportion $(1 - c)$. Therefore, the signal from allele B is decreased by a factor of $(1 - c)$. As a consequence, the B/A ratio is increased by a factor of $[(1 - c) / (1 + c)]$. For heterozygote SNPs, I_B/I_A is equal to 1. Using Equation 3, we obtain:

$$(1 - c)/(1 + c) = \tan\left(\mathrm{BAF}_{\mathrm{deldupl}} \times \pi/2\right) \qquad \text{[Equation 8]}$$

so that

$$c = \left(1 - \tan\left[\mathrm{BAF}_{\mathrm{deldupl}} \times \pi/2\right]\right)/\left(1 + \tan\left[\mathrm{BAF}_{\mathrm{deldupl}} \times \pi/2\right]\right) \qquad \text{[Equation 9]}$$

In order to determine which of the above three formulae should be used to compute c, *SOMATICS* assigns an a priori alteration type to each region in allelic imbalance, by use of the logR ratio intensity measurement. The mean logR ratio in that region is compared to a *reference* logR (see *SOMATICs* Code, Web Resources below). Regions with a significantly lower logR are assigned as deletions, those significantly higher as amplifications, and those not significantly different (at the 0.05 level) as deletion/duplication (also known as copy-neutral LOH or neutral allelic imbalance). On the basis of these a priori alteration-type assignments, the appropriate formula is used to determine c. Because significant variations of logR ratio can be noise-related artefacts, an a posteriori confirmation of the preliminary assignment is performed with c (see section iv of Appendix A). Any assignment of a deletion or an amplification that cannot be confirmed is considered as a deletion/duplication, and c is recomputed accordingly.

For any of these three somatic events, the band of heterozygous SNPs on the BAF scatter plot is split into two bands (Figure 1), one with an average BAF lower than 0.5 and the other with an average BAF higher than 0.5. The formulae above were defined for a BAF ratio lower than 0.5 (e.g., heterozygous SNPs in allelic imbalance with A being the "major" allele). To also include the information of heterozygous SNPs with BAFs higher than 0.5, these SNPs are converted into values lower than 0.5, "folding" as shown in Figure 2A (see section III of Appendix A).

Appendix C. Germline Genotype Inference from Cancer Samples

A reliable germline genotype can be inferred provided that the bands of heterozygous SNPs are distinct from the bands of homozygous SNPs. In chromosomal regions that are normal (regions with a single band of heterozygous SNPs), two fixed BAF thresholds are used to call the genotypes, corresponding to 0.25 and 0.75. SNPs with BAFs < 0.25 are called AA, those with BAFs between 0.25 and 0.75 are called AB, and those with BAFs > 0.75 are called BB.

In chromosomal regions that are in allelic imbalance, with two-band patterns that are obviously distinct from the bands of homozygous SNPs (red and green regions defined in Appendix A), two fixed BAF thresholds are used to call the genotypes corresponding to the 99.9th percentile of the reference BAF distributions of SNPs AA and BB.

In chromosomal regions that are in allelic imbalance, but with two-band patterns that are close to the bands of homozygous SNPs (blue region), *SOMATICs* models the distributions of homozygous and heterozygous SNPs. These model distributions are used to determine the number of true or false homozygous and heterozygous SNPs associated with each BAF threshold. A best-discrimination threshold can thus be identified.

The fitting of distribution models and the estimation of numbers of homozygous and heterozygous SNPs are performed differently in the situations of *distinct* (Figure S4A) and *overlapping* (Figure S4B) distributions. To discriminate between *distinct* and *overlapping* distributions, the first step is to assign an a priori genotype (homozygous or heterozygous) to the SNPs in this region. A liberal BAF-threshold value is used (95^{th} percentile of the *reference* homozygous distributions, Figure S4A). The distribution peak of a priori heterozygous SNPs is then determined as the BAF distribution class containing the highest number of SNPs, with a broad class-width definition (1/3 of the standard deviation of heterozygous *reference* SNPs, Figure S4A). The *reference* distribution of heterozygous SNPs is centered on this peak. If the 95^{th} percentile of this *reference* distribution centered on this peak is outside of the 95^{th} percentile of the homozygous *reference* distribution, the distributions of homozygous and heterozygous SNPs are considered *distinct* (Figure S2A), given that less than 5% of SNPs are expected to be misclassified. Otherwise, the distributions are considered as *overlapping* (Figure S4B).

In the case of *distinct* distributions, the distribution of a priori heterozygous SNPs is modeled with the *reference* distribution of heterozygous SNPs and the number of heterozygous SNPs is estimated as the number of a priori heterozygous SNPs. The distribution of homozygous SNPs is modeled with the *reference* distribution of homozygous SNPs, and the number of homozygous SNPs is estimated by counting of the SNPs with a BAF between 0 and the 75th first percentile of the *reference* distribution of AA SNPs (or between the 75th percentile of the *reference* distribution of BB SNPs and 1) and multiplying this number by 4/3 to reach 100%.

In the case of *overlapping* distributions, it is not possible to know whether SNPs in the overlapping region are homozygous or heterozygous. In addition, with BAF values so close to 0 or 1, the distribution of the SNPs belonging to the bands of the two-band patterns is tighter than the *reference* distribution of heterozygous SNPs. Therefore, the distribution of heterozygous SNPs is modeled with the largest 50% of the distribution of a priori heterozygous

SNPs, e.g., the distribution of SNPs with BAFs higher than the distribution peak, assuming that the distribution is symmetric. The number of heterozygous SNPs is estimated to be twice the number of these SNPs. The distribution of homozygous SNPs is modeled with the *reference* homozygous distribution, and the number of homozygous SNPs is estimated by counting of the SNPs with a BAF between 0 and the 50th percentile of the *reference* distribution of AA SNPs (or 1 and the 50th percentile of the *reference* distribution of BB SNPs) and multiplying this number by 2 to reach 100%.

The fitted distributions are used to generate cumulative distribution functions (CDFs) for the homozygous and heterozygous SNPs. The CDFs are combined for the estimated number of heterozygous and homozygous SNPs so that the numbers of false homozygous, true homozygous, false heterozygous, and true heterozygous SNPs can be estimated for each value of BAF (Figure S4C).

If one BAF threshold yields both true homozygote and true heterozygote rates greater than 99%, then this threshold is used for the discrimination between homozygous and heterozygous SNPs. If this criterion cannot be met, two distinct thresholds are defined, one for calling the homozygotes and the other for calling the heterozygotes, both with $\geq$ 99% true positives. SNPs with intermediate values are called "undetermined."

Supplemental Data

Supplemental data include four figures and can be found with this article online at http://www.ajhg.org/.

Acknowledgments

The authors thank Mohammed Orloff for critical review of drafts of the manuscript. This work is funded in part by a Bench–to-Bedside Translational Award from the National Institutes of Child Health and Human Development (NICHD), the National Institutes of Health, Office of Rare Diseases (NIH, ORD) (to C.A.S and C.E.), the Intramural Program of NICHD, NIH (to C.A.S.), and Cleveland Clinic Strategic Investment Funds to the Genomic Medicine Institute (to C.E.). G.A. is supported by clinical research fellowships from the Fondation de France and the Fédération Nationale des Centres de Lutte contre le Cancer. C.E. is a recipient of the Doris Duke Distinguished Clinical Scientist Award and is the Sondra J. and Stephen R. Hardis Chair of Cancer Genomic Medicine at the Cleveland Clinic.

Received: November 23, 2007
Revised: January 28, 2008
Accepted: January 29, 2008
Published online: March 20, 2008

Web Resources

The URLs for data presented herein are as follows:

Affymetix, www.affymetrix.com
dChip, http://biosun1.harvard.edu/complab/dchip/
Genomics Core Facility of Cleveland Clinic, www.lerner.ccf.org/services/gc
Illumina, www.illumina.com
R: R packages dip test and Adaptive Weights Smoothing (aws), http://www.r-project.org/
SOMATICs, http://www.lerner.ccf.org/gmi/igac

References

1. Dutt, A., and Beroukhim, R. (2007). Single nucleotide polymorphism array analysis of cancer. Curr. Opin. Oncol. *19*, 43–49.
2. LaFramboise, T., Harrington, D., and Weir, B.A. (2007). PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. Biostatistics *8*, 323–336.
3. Huang, J., Wei, W., Chen, J., Zhang, J., Liu, G., Di, X., Mei, R., Ishikawa, S., Aburatani, H., Jones, K.W., and Shapero, M.H. (2006). CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics *7*, 83.
4. Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M.E., et al. (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res. *16*, 1575–1584.
5. Lin, M., Wei, L.J., Sellers, W.R., Lieberfarb, M., Wong, W.H., and Li, C. (2004). dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics, in press.
6. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D.K., Kennedy, G.C., and Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. *65*, 6071–6079.
7. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. *35*, 2013–2025.
8. Yamamoto, G., Nannya, Y., Kato, M., Sanada, M., Levine, R.L., Kawamata, N., Hangaishi, A., Kurokawa, M., Chiba, S., Gilliland, D.G., Koeffler, H.P., and Ogawa, S. (2007). Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. Am. J. Hum. Genet. *81*, 114–126.
9. Gorringe, K.L., Jacobs, S., Thompson, E.R., Sridhar, A., Qiu, W., Choong, D.Y., and Campbell, I.G. (2007). High-resolution single nucleotide polymorphism array analysis of epithelial ovarian cancer reveals numerous microdeletions and amplifications. Clin. Cancer Res. *13*, 4731–4739.
10. Stark, M., and Hayward, N. (2007). Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. Cancer Res. *67*, 2632–2642.
11. Fitzgibbon, J., Iqbal, S., Davies, A., O'Shea, D., Carlotti, E., Chaplin, T., Matthews, J., Raghavan, M., Norton, A., Lister, T.A., and Young, B.D. (2007). Genome-wide detection of recurring sites of uniparental disomy in follicular and transformed follicular lymphoma. Leukemia *21*, 1514–1520.
12. Lips, E.H., de Graaf, E.J., Tollenaar, R.A., van Eijk, R., Oosting, J., Szuhai, K., Karsten, T., Nanya, Y., Ogawa, S., van de Velde,

C.J., et al. (2007). Single nucleotide polymorphism array analysis of chromosomal instability patterns discriminates rectal adenomas from carcinomas. The Journal of pathology. J. Pathol. *212*, 269–277.

13. Purdie, K.J., Lambert, S.R., Teh, M.T., Chaplin, T., Molloy, G., Raghavan, M., Kelsell, D.P., Leigh, I.M., Harwood, C.A., Proby, C.M., and Young, B.D. (2007). Allelic imbalances and microdeletions affecting the PTPRD gene in cutaneous squamous cell carcinomas detected using single nucleotide polymorphism microarray analysis. Genes Chromosomes Cancer *46*, 661–669.
14. Yu, Y., Baras, A.S., Shirasuna, K., Frierson, H.F., and Moskaluk, C.A. (2007). Concurrent loss of heterozygosity and copy number analysis in adenoid cystic carcinoma by SNP genotyping arrays. Lab. Invest. *87*, 430–439.
15. Kloth, J.N., Oosting, J., vanWezel, T., Szuhai, K., Knijnenburg, J., Gorter, A., Kenter, G.G., Fleuren, G.J., and Jordanova, E.S. (2007). Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex genetic alterations in cervical cancer. BMC Genomics *8*, 53.
16. Torring, N., Borre, M., Sorensen, K.D., Andersen, C.L., Wiuf, C., and Orntoff, T.F. (2007). Genome-wide analysis of allelic imbalance in prostate cancer using the Affymetrix 50K SNP mapping array. Br. J. Cancer *96*, 499–506.
17. Hu, N., Wang, C., Hu, Y., Yang, H.H., Kong, L.H., Lu, N., Su, H., Wang, Q.H., Goldstein, A.M., Buetow, K.H., et al. (2006). Genome-wide loss of heterozygosity and copy number alteration in esophageal squamous cell carcinoma using the Affymetrix GeneChip Mapping 10 K array. BMC Genomics *7*, 299.
18. Pfeifer, D., Pantic, M., Skatulla, I., Rawluk, J., Kreutz, C., Martens, U.M., Fisch, P., Timmer, J., and Veelken, H. (2007). Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. Blood *109*, 1202–1210.
19. Liu, W., Chang, B., Sauvageot, J., Dimitrov, L., Gielzak, M., Li, T., Yan, G., Sun, J., Sun, J., Adams, T.S., et al. (2006). Comprehensive assessment of DNA copy number alterations in human prostate cancers using Affymetrix 100K SNP mapping array. Genes Chromosomes Cancer *45*, 1018–1032.
20. Wang, Z.C., Buraimoh, A., Iglehart, J.D., and Richardson, A.L. (2006). Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. Breast Cancer Res. Treat. *97*, 301–309.
21. Andersen, C.C., Wiuf, C., Kruhoffer, M., Korsgaard, M., Laurberg, S., and Orntoff, T.F. (2007). Frequent occurrence of uniparental disomy in colorectal cancer. Carcinogenesis *28*, 38–48.
22. Gaasenbeek, M., Howarth, K., Rowan, A.J., Gorman, P.A., Jones, A., Chaplin, T., Liu, Y., Bicknell, D., Davison, E.J., Fiegler, H., et al. (2006). Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers. Cancer Res. *66*, 3471–3479.
23. Teh, M.T., Blaydon, D., Chaplin, T., Foot, N.J., Skoulakis, S., Raghavan, M., Harwood, C.A., Proby, C.M., Philpott, M.P., Young, B.D., and Kelsell, D.P. (2005). Genome-wide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event. Cancer Res. *65*, 8597–8603.
24. Irving, J.A., Bloodworth, L., Brown, N.P., Case, M.C., Hogarth, L.A., and Hall, A.G. (2005). Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. Cancer Res. *65*, 3053–3058.
25. Koed, K., Wiuf, C., Christensen, L.L., Wikman, F.P., Ziege, K., Molle, K., vanderMasse, H., and Orntoff, T.F. (2005). High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. Cancer Res. *65*, 34–45.
26. Calhoun, E.S., Hucl, T., Gallmeier, E., West, K.M., Arking, D.E., Maitra, A., Iacobuzio-Donahue, C.A., Chakravarti, A., Hruban, R.H., and Kern, S.E. (2006). Identifying allelic loss and homozygous deletions in pancreatic cancer without matched normals using high-density single-nucleotide polymorphism arrays. Cancer Res. *66*, 7920–7928.
27. Kotliarov, Y., Steed, M.E., Christopher, N., Walling, J., Su, Q., Center, A., Heiss, J., Rosenblum, M., Mikkelsen, T., Zenklusen, J.C., and Fine, H.A. (2006). High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. Cancer Res. *66*, 9428–9436.
28. Mullighan, C.G., Goorha, S., Radtke, I., Miller, C.B., Coustan-Smith, E., Dalton, J.D., Girtman, K., Mathew, S., Ma, J., Pounds, S.B., et al. (2007). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature *446*, 758–764.
29. Peiffer, D.A., Le, J.M., Steemers, F.J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C.A., Belmont, J., Cheung, S.W., Shen, R.M., Barker, D.L., and Gunderson, K.L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. *16*, 1136–1148.
30. Kurose, K., Gilley, K., Matsumoto, S., Watson, P.H., Zhou, X.P., and Eng, C. (2002). Frequent somatic mutations in *PTEN* and *TP53* are mutually exclusive in the stroma of breast carcinomas. Nat. Genet. *32*, 355–357.
31. Fukino, K., Shen, L., Patocs, A., Mutter, G.L., and Eng, C. (2007). Genomic instability within tumor stroma and clinicopathologic characteristics of sporadic primary invasive breast carcinomas. JAMA *297*, 2103–2111.
32. Weber, F., Xu, Y., Zhang, L., Patocs, A., Shen, L., Platzer, P., and Eng, C. (2007). Microenvironmental genomic alterations correlate with clinico-pathologic behavior in head and neck squamous cell carcinomas. JAMA *297*, 187–195.
33. Dumur, C.I., Dechsukhum, C., Ware, J.L., Coffield, S.S., Best, A.M., Wilkinson, D.S., Garrett, C.T., and Ferreira-Gonzalez, A. (2003). Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. Genomics *81*, 260–269.
34. Simon-Sanchez, J., Scholz, S., Fung, H.C., Matarin, M., Hernandez, D., Gibbs, J.R., Britton, A., de Vrieze, F.W., Peckham, E., Gwinn-Hardy, K., et al. (2007). Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum. Mol. Genet. *16*, 1–14.
35. Polzehl, J., and Spokoiny, S. (2000). Adaptative weights smoothing with applications to image restoration. J. Roy. Statist. Soc. Ser. B. Methodological *62*, 335–354.
36. R Development Core Team (2005). R: A language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing). http://www.R-project.org.

37. Maechler, M. (2007) Hartigan's dip test statistic for unimodality–corrected code R package 'diptest', version 025–1. http://cran.r-project.org/web/packages/diptest/index.html.
38. Oliphant, A., Barker, D.L., Stuelpnagel, J.R., and Chee, M.S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. BioTechniques S56–S58, S60–S61.
39. Pasini, B., et al. (2008). Clinical and molecular genetics of patients with the Carney-Stratakis syndrome and germline mutations in the genes encoding succinate dehydrogenase subunits (*SDHB, SDHC, SDHD*). Eur. J. Hum. Genet. *16*, 79–88.
40. Hartigan, J.A., and Hartigan, P.M. (1985). The dip test of unimodality. Ann. Stat. *13*, 70–84.